DOCUMENT RESUME

ED 283 876 TM 870 415

AUTHOR Cohen, Peter A.
TITLE A Critical Analysis and Reanalysis of the Multisection Validity Meta-Analysis.
PUB DATE Apr 87
NOTE 28p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Correlation; *Effect Size; Field Studies; Higher Education; *Meta Analysis; Rating Scales; Research Methodology; Research Problems; *Student Evaluation of Teacher Performance; *Teacher Characteristics; Validity

ABSTRACT

This paper focuses on a critical analysis and reanalysis of two earlier multisection validity meta-analyses of student ratings of instruction. The reanalysis is based on a sharpening of criteria for: (1) including studies in the meta-analysis; (2) defining key study characteristics that influence results; and (3) calculating the correlational effect size (between instructor rating and student achievement). A subset of 41 well-designed studies and a further subset of 25 high-quality studies were identified, and correlational effect sizes were found comparable to those for the entire sample (for well-designed studies) or slightly smaller (for high-quality studies). The results are seen as supporting the validity of student ratings, especially global ratings of the instructor, and certain rated dimensions such as skill, rapport, structure, interaction, evaluation, and students' self-rating of their learning. Despite overall support, a number of studies showed negative or near-zero effect sizes, but methodological problems may have played a role in producing negative correlations in two studies. (LPG)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED283876

A Critical Analysis and Reanalysis of the Multisection Validity Meta-analysis

Peter A. Cohen, Ph.D.

Division of Instructional Development
The University of Texas
Health Science Center at San Antonio
7703 Floyd Curl Drive
San Antonio, Texas 78284

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

P Cohen

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Presented at the Annual Meeting of the
American Educational Research Association

Washington, D.C. 1987

[Please do not cite or quote without permission]

TM 870 415

BEST COPY AVAILABLE

A Critical Analysis and Reanalysis of the Multisection Validity Meta-analysis

This symposium brings together a number of researchers who have conducted meta-analyses related to student ratings of instruction. The theme of the symposium centers around the processes by which decisions are made and conclusions drawn in a meta-analysis. We will share our collective experiences in conducting large scale quantitative reviews, discuss the advantages and disadvantages of such an approach, and recommend future courses of action.

A major purpose of meta-analysis is to explain conflicting results in a body of studies. It seems a particularly useful approach for synthesizing research results in controversial research areas. The student rating validity literature is one such area, and indeed several independent meta-analyses have been conducted on this literature. Unfortunately, rather than bringing research findings into sharper focus, these meta-analyses of student rating validity have come to strikingly different conclusions. What we now have are conflicting results in a body of meta-analyses.

Abrami (1986) attempted to explain different conclusions concerning student rating validity among six meta-analyses conducted in this area. He found that the different conclusions reached by these reviewers related to differences at each step in

the review process -- statement of the problem, identification of studies to be included in the analysis, coding of studies, and data analysis. While Abrami pointed out that these inconsistencies were not a weakness of meta-analysis, his findings indicate the need to critically examine procedures and conclusions of quantitative reviews. In this light, my presentation will focus on a critical analysis and reanalysis of two earlier multisection validity meta-analyses (Cohen, 1981; 1986).

Decisions Made in Earlier Meta-analyses

At least a dozen narrative reviews and three independent statistical, or meta-analytic reviews, have drawn conclusions about the relationship between student ratings and student learning. These reviews summarize a large body of what has come to be known as multisection validity studies. Although there are inherent weaknesses in the generalizability of the design, multisection studies may be the best evidence we have to address the validity question. Commenting on the multisection validity approach Marsh (1984) states, "Perhaps more than any other type of study, the credibility of student ratings has rested on this paradigm."

In 1981, I reviewed the multisection validity literature using a meta-analytic approach. Along with Philip Abrami and his colleagues at Concordia University, I updated this meta-analysis in 1986. The primary purpose of these meta-analyses was to

provide an accurate reflection of existing literature. To a lesser extent, I was interested in estimating true correlational effect sizes between student ratings and student learning. Decisions I made in these meta-analyses related to these purposes. For this presentation, I would like to examine decisions made in the earlier meta-analyses with particular regard to inclusion criteria, study characteristic definitions, and effect size calculations.

Inclusion criteria. In the 1981 and 1986 meta-analyses, a study had to meet four general criteria to be included in the final sample. First, the study had to provide data from actual college or university classes. In other words, the study had to be a "field" study rather than an experimental analogue of teaching. Second, the unit of analysis in the study had to be the class (instructor) rather than the individual student. Third, the study needed to use a multisection format with a common achievement measure assessing learning for all sections. And fourth, the study had to provide data from which a ratings/achievement correlation could be calculated. Descriptive reports that did not provide enough data for the calculation of a correlational effect size were excluded.

These inclusion criteria were developed so that studies in the final sample would be relatively homogeneous with respect to the basic multisection validity paradigm. Because the primary purpose of the two earlier meta-analyses was to statistically

summarize the existing literature, I felt it was important to include all studies that met these criteria regardless of study quality.

Study characteristic definitions. A total of 20 variables were used to describe methodologies, settings, and publication features of the studies. Some of these study characteristics had been discussed in relation to the student rating validity literature by other researchers and reviewers, such as assignment of students to teachers, control for initial student ability, instructor experience, rating instrument bias, and number of sections in the course. Other characteristics -- prior knowledge of the instructor and the time at which the ratings were administered, for example -- I identified as being of potential importance after reading through studies collected for the initial sample. Finally, I included study characteristics such as study quality, type of institution, and publication features because they had correlated with outcomes in other meta-analyses.

Some of these study characteristics related to the overall instructor rating/learning correlational effect size. A finding of interest in the earlier meta-analyses was that when ratings were administered after students knew their final grades, the ratings/learning correlations were significantly higher. Here, a potential source of bias -- Are good teachers receiving high ratings or are students rewarding teachers who give high grades? -- was empirically demonstrated to affect the correlational

effect size. Another study characteristic that I found related to correlational effect size was control for teacher bias in evaluating achievement. Correlations between overall instructor ratings and learning were much smaller in studies where students' own instructors, as opposed to an external evaluator, graded achievement examinations. Finally, the number of sections in a multisection course affected the size of the correlational effect. Correlations were higher and more variable in studies with small numbers of sections. Studies using 10 or more sections seemed to more accurately reflect the true effect.

The empirical findings involving these methodological study features suggest that study quality, and perhaps inclusion criteria, need to be redefined. In the early analyses, study quality was operationally defined by summing scores from the other coded methodological variables. Some of those methodological characteristics probably affected overall study quality greatly, while other characteristics related little to study quality. My current feeling is that we can more accurately assess study quality by defining a priori criteria for determining methodological biases.

Effect size calculations. In the earlier meta-analyses I decided that it was important to generate a "summary instructor effect" for each study. To accomplish this, I had to average over different rating dimensions (e.g., skill, structure, rapport, interaction) to obtain the overall effect size for some

studies. In retrospect, this approach was suspect for two reasons. First, student ratings are multidimensional in nature. Thus, by using different combinations of dimensional ratings, or a single global rating, I was not obtaining a consistent or pure measure of an overall instructor effect among the set of studies. Second, when we talk about the validity of an overall instructor rating, we almost always refer to a single global item on a rating form. Averaging specific items or dimensional ratings has little practical application for determining the overall effectiveness of the instructor.

Another problem concerning the calculation of effect sizes had to do with categorization of dimensional ratings. Factor-analytic research consistently has identified a number of common dimensions of teaching, among them skill, rapport, structure, difficulty, interaction, feedback, and evaluation. Unfortunately, researchers tend to throw this dimensional nomenclature around rather loosely. The actual items included on a rating form may be disparate with the name of the dimension which supposedly describes them. In reviewing the actual items used in the set of studies for the reanalysis of these data, I discovered a number of discrepancies which did not surface in the earlier meta-analyses.

Reanalysis of Overall and Dimensional Ratings

For the current reanalysis of the multisection validity data, I reviewed the original categorizations for overall and

dimensional student ratings. I distinguished between global (G) ratings and scaled (S) ratings for the overall course and overall instructor ratings. Global ratings were single rating items assessing the overall effectiveness of the course or instructor [e.g., "This is an excellent course," "Rate your instructor"]. Scaled ratings were summated overall rating items or a scaled factor on overall effectiveness. I did not consider averages of dimensional ratings (e.g., skill, rapport, structure) to be a pure enough index of overall effects. Table 1 shows overall instructor rating/learning correlational effect sizes from the earlier analysis (Cohen, 1986) and those generated for the current reanalysis. Thirteen of the 1986 effect sizes were not included in the reanalysis because they were not either global or scaled overall ratings. One additional report (Greenwood et al., 1976) was dropped from the analysis because it was the same study as Hazelton (1980). Thus, the reanalysis included effect sizes from 59 multisection courses. In five of these studies, 1986 effect sizes were altered for the reanalysis. In two cases, updated reports provided different effect sizes; in one case lecture section results were used in place of laboratory results, and in one report (two multisection courses) the correlational effect sizes had been incorrectly calculated from T-tests.

Insert Table 1 about here

Table 2 presents mean rating/learning correlations for the

overall instructor and overall course ratings. The mean correlational effect sizes computed for the reanalysis are very similar to those found in the earlier analysis. One finding of interest here was the difference in effect size between global and scaled overall instructor ratings. The mean validity coefficient was higher for global ratings, although this difference was not statistically significant, t (57) = 1.05, $p > .05$.

Insert Table 2 about here

Table 3 presents mean rating/learning correlations for the dimensional ratings. Mean correlational effect sizes for the current reanalysis were very similar to those found in the 1986 study. I further broke down these dimensional ratings as to how they were actually derived. Ratings could be based on a coherent dimension or scale (usually derived from a factor-analytic procedure), a summated average of specific rating items addressing the dimension category, or a single rating item falling under the dimension category. Most studies provided data for either dimension or single item ratings. Although no solid pattern emerges, validity coefficients based on dimensions seem to be higher than single item ratings for the rapport, interaction, evaluation, and learning dimensions.

Insert Table 3 about here

Study Quality and Effect Size

Reviewers of the multisection validity literature have discussed criteria for well-designed studies (Abrami, 1986; Cohen, 1983, 1986; Crooks and Smock, 1980; Dowell and Neal, 1982). These criteria include the following:

1. There is equivalence of student ability among sections.
2. Ratings are administered before students know their final grade.
3. Teachers are primarily responsible for instruction.
4. Students have no prior knowledge of the instructor before enrolling in the section.
5. A single global rating item or internally consistent scale is used.
6. A sufficient number of sections is used to assess the relationship between ratings and learning.

To determine the effect of study quality on effect size, I identified studies that met certain criteria. First of all, studies had to report evidence of equivalence of student ability. This could be shown either through random assignment of students to sections, evidence of initial section equivalence regarding student ability, or through statistical control of initial

student ability. Also, teachers had to be primarily responsible for instruction and students had to rate instructors before they found out their final grades in order to reduce possible grade/rating bias. Based on these criteria 41 of the 73 studies were included in what I called the "well-designed studies" sample. Correlational effect sizes for this subset of studies are presented in Table 4.

Insert Table 4 about here

In going one step further in the analysis of study quality, I defined a subset of "high quality" studies that met the following two additional criteria: a) There were at least 10 sections in the multi-section course, and b) The study used a global rating item for the overall instructor and overall course ratings and/or used an internally consisent factor for the dimensional ratings. Twenty-five studies met these criteria. Correlational effect sizes for this set of "high quality" studies are presented in Table 5.

Insert Table 5 about here

Correlational effect sizes for the well-designed studies were comparable to the effect sizes for the entire sample. The mean correlations for the selected high quality studies, however,

were somewhat smaller in magnitude. Still, correlations between ratings and learning for the overall rating items and many of the dimensional ratings were strong enough to support the validity of student ratings.

Studies Which Do Not Support the Validity of Student Ratings

The overall results of the earlier meta-analyses and the current reanalysis support the validity of student ratings. There were a number of studies in the sample, however, that produced negative or near zero correlational effect sizes. These studies are listed in Table 6. Of the 11 multisection courses shown in the table, only two (Doyle and Crichton; Sullivan and Skanes-Psychology) were included in the set of high quality studies. The two studies with large negative correlational effect sizes (Bendig; Rodin and Rodin) each had numerous methodological problems and potential study biases. We can be fairly confident that the outcomes from these studies do not accurately reflect the state of student rating validity.

Insert Table 6 about here

Discussion and Conclusions

The multisection validity literature supports the validity of student ratings. If we look at only the high quality studies in this area, results correspond quite well with those found in the earlier meta-analyses. As one who is called upon to advise administrators on the use of student ratings, I am confident that global ratings of the instructor and course, and certain rating dimensions such as skill, rapport, structure, interaction, evaluation, and students' self-rating of their learning can be used effectively as an integral component of a teaching evaluation system.

Although my views on the validity of student ratings have not changed, I have become more of a skeptic about the usefulness of meta-analysis to resolve issues in this area. In 1980, I had hopes that the original multi-section validity meta-analysis would resolve some of the controversy regarding differing conclusions drawn from this literature. It did not. I am of the impression that the most cited studies carry an undue weight and that a relatively small number of studies influence greatly generalizations of reviewers. I also believe that studies are selected by reviewers not because they are representative of the literature, but rather to support a particular point of view regarding student rating validity. For example, the two most cited multisection validity studies are Rodin and Rodin (1972),

showing a large negative correlation, and Frey (1973), showing a large positive correlation (cf, Cohen, 1983). These two studies are not, however, representative of the area in terms of either effect size or study quality. In point of fact, both of these studies have serious methodological flaws. Meta-analytic research supposedly serves to temper reviewers' subjectivity. In the area of student rating validity, I don't think it has accomplished this.

Perhaps I have put an inordinate amount of faith in the objectivity of meta-analysis. In conducting meta-analysis there are many decision points which basically involve subjective judgment on the part of the researcher. Meta-analysts have been good, in general, at clearly specifying those decisions. Outlining study inclusion criteria, for example, has been a substantial improvement over traditional narrative review techniques. Abrami (1986) was able to conclude that differences in conclusions reached by meta-analytic reviewers related to differences at each stage of the review process only because these steps were fairly clearly indicated in each of the reviews. In this sense, meta-analysis has helped us evaluate the quality of quantitative and qualitative reviews of student rating validity.

Even if we can maintain a fair amount of objectivity, meta-analysis has limitations. One problem involves the lack of control a meta-analyst has regarding reported results of primary

research. A meta-analyst is limited by what data the researcher chooses to present. In some cases, reported results may be misleading or even incorrect. There are also many potential errors a meta-analyst can make, for instance, in the coding of studies and in the calculation of effect sizes. I made some of these errors in the earlier analyses, but I was fortunate to have colleagues critically examine my initial decisions. As it turned out, the outcomes were not affected much by these decisions.

In my mind, meta-analytic research of the student rating validity literature clearly supports the usefulness of student ratings -- particularly global ratings -- as a source of information on teaching effectiveness. But in a sense I am frustrated that this method of research has had what I consider to be low impact in the area. Meta-analysis as a method of reviewing the literature has been popular with reviewers, but I do not see results of meta-analyses being given any more weight than traditional narrative reviews. Some researchers claim that meta-analysis cannot, by itself, give a representative picture of a research area (e.g., Slavin, 1986). As we concentrate heavily on the statistical aspects of meta-analysis, I am afraid we are losing touch with the consumers of quantitative reviews. Regarding the student rating validity literature, it is time to concentrate our efforts on helping practitioners extract meaning from meta-analytic research. That should be our challenge and future direction.

References

Abrami, P.C. (1986, April). What quantitative reviews say about the validity of student ratings. In R.P. Perry (chair), Student ratings of instruction: A new look at multisection validity studies. Symposium conducted at the annual meeting of the American Educational Research Association, San Francisco, CA.

Cohen, P.A. (1981). Student ratings of instruction and student achievement. A meta-anaylsis of multisection validity studies. Review of Educational Research, 51, 281-309.

Cohen, P.A. (1983). Comment on a selective review of the validity of student ratings of teaching. Journal of Higher Education, 54, 448-458.

Cohen, P.A. (1986, April). An updated and expanded meta-analysis of multisection student rating validity studies. In R.P. Perry (chair), Student ratings of instruction: A new look at multisection validity studies. Symposium conducted at the annual meeting of the American Educational Research Association, San Francisco, CA.

Crooks, T.J., and Smock, H.R. (1980). Student ratings of instructors related to student achievement. Urbana, Ill.: Office of Instructional Resources, University of Illinois.

Dowell, D.A., and Neal, J.A. (1982). A selective review of the validity of student ratings of teaching. Journal of Higher Education, 53, 51-62.

Frey, P.W. (1973). Student ratings of teaching: Validity of several rating factors. Science, 182, 83-85.

Greenwood, G.E., et al. (1976). A study of the validity of four types of student ratings of college teaching assessed on a criterion of student achievement gains. Research in Higher Education, 5, 171-178.

Hazelton, A.E. (1980). A study of the validity of student ratings of college teaching assessed on a criterion of student achievement in a first course in calculus. Unpublished doctoral dissertation, The University of Florida.

Marsh, H.W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. Journal of Educational Psychology, 76, 707-754.

Rodin, M., and Rodin, B. (1972). Student evaluations of teachers. Science, 177, 1,164-1,166.

Slavin, R.E. (1986). Best-evidence synthesis: An alternative to meta-analytic and traditional reviews. Educational Researcher, 15(9), 5-11.

Table 1

Overall Instructor Rating/Learning Correlational Effect Sizes

Study[1]	Finding #	ES 1986	ES 1987	Reason for Change in Effect Size Calculation
Bendig - 1953a	01	.90	.90(S)[2]	
Bendig - 1953b	02	-.80	-.80(S)	
Benton & Scott - 1976	03	.17	-	Average of dimensional ratings
Bolton et al. - 1979	04	.68	.68(G)	
Borg & Hamilton - 1956	05	.19	.19(S)	
Braskamp et al. - 1979	06	.17	.17(G)	
	07	.48	.48(G)	
Bryson - 1974	08	.55	.55(G)	
Centra - 1979	09	.41	.41(G)	
	10	.58	.58(G)	
	11	.87	.87(G)	
	12	.23	.23(G)	
	13	.64	.64(G)	
	14	.61	.61(G)	
	15	.60	.60(G)	
Chase & Keene - 1979	16	.38	.24(G)	Miscalculation of ES from T-test
	17	.93	.96(G)	
Cohen & Berger - 1970	18	.48	.48(S)	
Costin - 1978	19	.56	.56(S)	
	20	.46	.46(S)	
	21	.41	.41(S)	
	22	.52	.52(S)	
Crooks & Smock - 1980	23	.44	.49(G)	Updated report (1980). Authors deleted one section because of change in instructors
Doyle & Crichton - 1978	24	-.04	-.04(G)	
Doyle & Whitely - 1978	25	.49	.49(G)	
Ellis & Rickard - 1977	26	.58	.59(G)	Use of global rating rather than scaled rating
Elliott - 1950	27	.33	.35(G)	Use of recitation results rather than laboratory results
Endo & Della-Piana - 1976	28	-.15	-.15(G)	
Frey - 1973	29	.91	-	Dimensional rating (skill)
	30	.60	-	
Frey - 1976	31	.79	-	Dimensional rating (skill)
Frey et al. - 1975	32	.81	-	
	33	.18	-	
Frey et al. - 1975	34	.74	-	
Greenwood et al. - 1976	35	-.11	-	Same study as Hazelton - 1980
Grush & Costin - 1975	36	.45	.45(S)	

Table 1 (continued)

Study	Finding #	ES 1986	ES 1987	Reason for Change in Effect Size Calculation
Hoffman - 1978	37	.27	-	Average of dimensional ratings
	38	.27	-	
Marsh et al. - 1975	39	.42	.42(G)	
McKeachie et al. - 1971	40	.26	-	Average of specific rating items
Mintzes - 1977	41	.38	-	Average of specific rating items
Morsh et al. - 1956	42	.40	.40(G)	
Murdock - 1969	43	.77	.77(G)	
Overall & Marsh - 1980	44	.41	.38(G)	Updated report - One additional section included in analysis
Rankin et al. - 1965	45	-	-	
Remmers et al. - 1949	46	.49	.49(G)	
Reynolds & Hansvick - 1978	47	.20	.20(G)	
Rodin & Rodin - 1972	48	-.75	-.75(G)	
Rubinstein & Mitchell - 1970	49	.14	.14(G)	
Solomon et al. - 1964	50	.46	.46(G)	
Sorge & Kline - 1973	51	.39	-	Average of dimensional ratings
Spencer & Dick - 1964	52	.88	.88(G)	
Sullivan & Skanes - 1974	53	.51	.51(G)	
	54	.01	.01(G)	
	55	.53	.53(G)	
	56	.57	.57(G)	
	57	.33	.33(G)	
	58	.34	.34(G)	
	59	.48	.48(G)	
	60	.55	.55(G)	
	61	.08	.08(G)	
	62	.42	.42(G)	
	63	-.28	-.28(G)	
Turner & Thompson - 1974	64	.08	.08(S)	
	65	-.15	-.15(S)	
Wherry - 1952	66	.16	.16(S)	
Whitely & Doyle - 1979	67	.80	.80(G)	
	68	-.11	-.11(G)	
Hazelton - 1980	69	.10	.10(G)	
McKeachie et al. - 1978	70	.22	.22(G)	
Morgan & Vaschè - 1978	71	.97	-	Average of dimensional ratings
Murray - 1983	72	.30	.30(G)	
Orpen - 1980	73	.52	.52(S)	
Soper - 1973	74	-.17	-.17(S)	

[1]Studies are referenced in Cohen, 1981 and Cohen, 1986

[2](G) = Global rating
(S) = Scaled rating

Table 2

Mean Overall Rating/Learning Correlational Effect Sizes

Rating Dimension	N	Mean Correlation	95% Confidence Interval
1986 Overall Instructor	73	.44	.34, .53
1987 Overall Instructor	59	.40	.30, .50
Global	45	.43	.31, .54
Scaled	14	.30	-.03, .56
1986 Overall Course	23	.46	.28, .62
1987 Overall Course	21	.48	.27, .65
Global	17	.45	.19, .65
Scaled	4	.60	-.04, .82

Table 3

Mean Dimension Rating/Learning Correlational Effect Sizes

	Mean Correlation				
Rating Dimension	1986	1987 Total	1987 Dimension	1987 Summated	1987 Single Item
Skill	.53 (43)[1]	.51 (44)	.51 (26)	.47 (5)	.52 (13)
Rapport	.31 (31)	.30 (35)	.34 (25)	.22 (1)	.20 (9)
Structure	.52 (30)	.55 (29)	.55 (23)	.40 (1)	.55 (5)
Difficulty	.01 (26)	.00 (25)	.02 (21)	(0)	-.09 (4)
Interaction	.42 (16)	.45 (20)	.50 (14)	(0)	.31 (6)
Feedback	.29 (6)	.29 (7)	.28 (2)	.34 (1)	.29 (3)
Evaluation	.23 (26)	.23 (25)	.33 (17)	(0)	-.01 (8)
Interest/Motivation	N/A[2]	.26 (12)	.19 (9)	.46 (1)	.42 (2)
Learning	.47 (17)	.46 (17)	.59 (9)	.55 (1)	.27 (7)

[1] ( ) Number of multisection courses

[2] Interest/Motivation dimension not included in earlier meta-analyses

Table 4

Correlational Rating/Learning Effect Sizes for
41 "Well-Designed" Studies

Rating Dimension	N	Mean Correlation	95% Confidence Interval
Overall Instructor	35	.45	.36, .54
Overall Course	16	.49	.23, .68
Skill	24	.50	.34, .64
Rapport	22	.32	.19, .45
Structure	17	.55	.24, .77
Difficulty	14	-.04	-.28, .21
Interaction	11	.52	.00, .82
Feedback	5	.28	.17, .39
Evaluation	15	.30	.10, .48
Interest/Motivation	3	.15	-.11, .39
Learning	11	.39	.16, .58

Table 5

Mean Rating/Learning Effect Sizes for
25 "High Quality" Studies

Rating Dimension	N	Mean Correlation	95% Confidence Interval
Overall Instructor	21	.41	.33, .49
Overall Course	9	.32	.17, .46
Skill	9	.40	.24, .53
Rapport	9	.30	.08, .48
Structure	9	.29	.06, .50
Difficulty	8	-.07	-.39, .27
Interaction	6	.38	.08, .61
Feedback	2	.28	-.62, .86
Evaluation	7	.36	.27, .45
Interest/Motivation	2	.14	-.53, .70
Learning	9	.42	.15, .62

Table 6

Studies with Negative or Near Zero Effect Sizes

Study	Instructor Experience	Number of Sections	Subject Matter	Correlational Effect Size	Potential Study Biases
Bendig-1953b	TA/FAC	5	Psychology	-.80	No ability control; Teachers graded exams for own section; Letter grade as achievement criterion; Limited to sophomore students only
Doyle & Crichton-1978	TA	12	Communications	-.04	
Endo & Della-Piana-1976	FAC	5	Mathematics	-.15	
Rodin & Rodin-1972	TA	12	Mathematics	-.75	Teachers graded achievement measure; Achievement criterion confounding; Teachers (TAs) not primarily responsible for instruction; Students had prior knowledge of instructor
Sullivan & Skanes-1974	TA	13	Psychology	.01	
	FAC	8	Chemistry	.08	
	FAC	6	Biology	-.28	
Turner & Thompson-1974	FAC	24	French	.08	No ability controls; Teachers graded exams for own section
	FAC	16	French	-.15	
Whitely & Doyle-1979	TA	11	Mathematics	-.11	Teachers not primarily responsible for instruction
Soper-1973	TA	14	Economics	-.17	No ability control; Teachers not primarily responsible for instruction